CUHK & ETHZ & SIAT Submission to ActivityNet Challenge 2016

Yuanjun Xiong[1], Limin Wang[2], Zhe Wang[3], Bowen Zhang[3], Hang Song[1], Wei Li[1], Dahua Lin[1], Yu Qiao[3], Luc Van Gool[2] and Xiaoou Tang[1]

[1]Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong
[2]Computer Vision Lab, ETH Zurich, Switzerland
[3]Shenzhen Institutes of Advanced Technology, CAS, China

Abstract

This paper presents the method that underlies our submission to the untrimmed video classification task of ActivityNet Challenge 2016. We follow the basic pipeline of very deep two-stream CNN [16] and further raise the performance via a number of other techniques. Specifically, we use the latest deep model architecture, e.g. ResNet and Inception V3 and introduce a new aggregation scheme (top-k and attention-weighted pooling). Additionally, we incorporate the audio as a complementary channel, extracting relevant information via a CNN applied to the spectrograms. With these techniques, we derive an ensemble of deep models, which, together, attains a very high classification accuracy (mAP 93.23%) on the testing set and secured the first place in the challenge.

1. Introduction

In the past several years, the advance in deep learning techniques has given rise to a new wave of efforts towards vision-based action understanding. A number of deep learning based frameworks, including two-stream CNNs [8], 3D CNNs (C3D) [12], and Trajectory-pooled Deep convolutional Descriptors (TDD) [14], have been developed, which significantly pushed forward the state-of-the-art [13, 15]. Such improvement on performance, to a large extent, is owning to both the modeling capacity of deep architectures and more effective learning strategies.

However, it is worth noting that previous efforts focus mainly on the analysis of short video clips. These clips are typically extracted from longer videos such that they only contain the portions of frames that truly capture the actions of interest. Obviously, preparation of such data is a laborious procedure. Action recognition from *untrimmed videos*, a problem that is more pertinent to real-world demands, is drawing increasing attention from the community. While substantially reducing the efforts needed in manual annotation, this task on the other hand presents a new challenge to the recognition system – a significant (or even dominant) fraction of a given video is irrelevant to the action of interest.

Driven by the ActivityNet benchmark [1], we develop an integrated approach to recognizing actions from untrimmed videos[1]. Our approach follows the framework of very deep two stream CNNs presented in our earlier paper [16], which allows both appearance and motion patterns to be effectively fused and introduces various techniques to improve the training procedure, *e.g.* temporal pre-training, and scale jittering augmentation. On top of this framework, we develop several new techniques to further improve the recognition accuracy. While visual analysis plays a primary role in this task, we notice that the audio channels that come with these videos provide complementary information. To exploit such information, we develop a deep network called Audio CNN to derive complementary features from the spectrograms.

Combining both the visual and acoustic models, we attain a high recognition accuracy (mAP 93.23% on testing set). We want to emphasize that this performance is obtained only using the training data provided by the ActivityNet benchmark except using CNNs pre-trained on ILSVRC12 data for initialization – no additional data or annotations are used throughout both the training and testing procedures.

The rest of this paper is organized as follows. Section 2 presents our approach in detail, Section 3 reports our results under a variety of settings, finally Section 4 concludes this work.

---

[1]Codes and models are available at `https://github.com/yjxiong/anet2016-cuhk`

Table 1. Performance of different network architectures on ActivityNet v1.3 validation set. Performance is measured by per-class mean average precision (mAP) and top-3 prediction accuracy. We use the variant "basic+a" in training these models.

Settings	Spatial Nets			Temporal Nets		
	BN-Inception	Inception V3	ResNet	BN-Inception	Inception V3	ResNet
mAP	79.7%	83.3%	83.3%	63.3%	64.3%	-
Top-3 Acc.	89.6%	91.5%	91.6%	77.0%	77.9%	-

2. Our Approach

Our approach to untrimmed video classification comprises two complementary components: visual and acoustic modeling. The visual analysis, which combines a variety of techniques, plays a primary role in this framework, while the acoustic model exploits complementary information from the audio channels to further improve the performance. Next, we present these components respectively in Section 2.1 and 2.2.

2.1. Visual Analysis System

Our visual analysis component works as follows: it samples multiple snippets from a given video, makes snippet-wise predictions using very deep two-stream CNNs, and finally aggregates the predictions via different strategies such as top-k and attention-weighted pooling.

Snippet-wise Predictor Deep convolutional neural networks (CNN) which learns from multiple modality of input data has been used extensively in visual recognition tasks [9, 18, 19, 2] and achieved superiority over models using a single modality. The snippet-wise predictor in our approach is a realization of the very deep two-stream CNN framework [16] which consists appearance and motion modeling parts. In this work, we adopt the recently proposed network architectures such as **ResNet** [3] and **Inception V3** [10] to improve the capacity of the frame-wise predictor.

During training of the snippet-wise predictor, the techniques introduced in [16], such as scale jittering and stronger dropout, is also applied to the these architectures. To further boost the performance, we experimented with the idea to sample several snippets from one input video to jointly train the CNNs by averaging the per-snippet prediction. This also enables us to apply more advanced aggregation techniques into the training process.

Video-level Classification To obtain video-level classification results, we use the following strategy: the snippet-wise predictor is first applied to an input video snippet with a 1FPS sampling rate, then an aggregation module will combine the snippet-wise class scores into the final prediction. We experimented with several advanced strategies for combing snippet-wise scores of the appearance nets.

Table 2. Performance comparison of the appearance modeling CNN variants on the validation set of ActivityNet v1.3. Here we analyze their performance using the Inception V3 [10] architecture. In the table, "basic" refers to the baseline approach in [16], "a" refers to models trained with multiple snippets from one video, "b" refers to models equipped with advanced aggregation strategies.

Variants	mAp	Top-3 Acc.
basic	82.9%	91.0%
basic+a	83.3%	91.5%
basic+ab	84.2%	92.1%
Ensemble	85.9%	92.9%

Table 3. Performance of different components in the visual analysis system on the validation set. Here, "Appearance CNN" refers to the appearance modeling part. "Motion CNN" refers to the motion modeling part. "Combined CNN" refers to the results by combining both appearance and motion modeling parts. "Visual All" refers to the results by further combining scores from other methods such as IDT [13, 6] and TDD [14].

Variants	mAp	Top-3 Acc.
Appearance CNN	85.9%	92.9%
Motion CNN	68.3%	80.2%
Combined CNN	89.7%	95.0%
Visual All	90.4%	95.2%

These include top-k pooling and attention weighted pooling. These strategies, when used in both training and testing, produced models that are complementary to each other and thus form effective components in the final ensemble.

2.2. Acoustic Analysis System

Audio signals in a video carry important cues for recognizing some action classes. To harness the information in this aspect, we combine the standard MFCC [5] representations with audio-based CNNs [11, 17] to form the acoustic modeling system.

MFCC Mel Frequency Cepstral Coefficients (MFCC) [5] is a powerful feature descriptor used in automatic speech recognition system. In our approach, we extract MFCC features from companioned audios of the videos in the dataset, and train SVMs on descriptors aggregated with Fisher Vector [7]

Table 4. Performance of acoustic models on ActivityNet v1.3 validation set. Performance is measured by per-class mean average precision (mAP) and top-3 prediction accuracy. Here, "Gray" refers to the models trained with grayscale inputs. "MS" refers to the model trained with multiple time scales.

Methods	mAP	Top-3 Acc.
MFCC (FV+SVM)	14.2%	26.1%
Audio CNN	8.0%	17.1%
Audio CNN Gray	9.3%	19.3%
Audio CNN Gray+MS	10.3%	20.7%
Audio Ensemble	15.2%	29.1%

Table 5. Performance of fusion models on ActivityNet v1.3. Performance is measured by per-class mean average precision (mAP) and top-3 prediction accuracy. In "Visual + Audio" setting, we combine the visual and acoustic modeling system. On the testing set, we present the results of "Final Ensemble" where all components trained on training plus validation data are combined.

Validation Set	mAp	Top-3 Acc.
Visual	90.4%	95.2%
Audio	15.2%	29.1%
Visual + Audio	90.9%	95.6%
Testing Set	mAP	Top-3 Acc.
Visual CNN (Single)	91.2%	95.6%
Final Ensemble	93.2%	96.4%

Audio CNN The basic idea of Audio CNN works is to apply CNNs on spectrograms, or time-frequency-response maps, of audio signals. In this work, we propose to directly use the *grayscale* time-frequency map image to train the audio CNN. Then the audio CNN can be initialized by the same technique used on the temporal networks in [16]. It is also known that learning from multiple time scales help in acoustic models [20]. In this sense, we propose to stack multiple spectrograms with varying window size as the input to the audio CNN.

3. Experiments

We train our models on the official training set of ActivityNet v1.3 dataset [1]. There are $10,024$ videos for training, enclosing 15410 activity instances from 200 activity classes. The validation set contains 4926 videos and 7654 activity instances. We study the performance of our approach on this validation set. The final testing set comprises 5044 videos and is not annotated with any activity instance. We report the performance of our proposed models on this set according to the feedback of the test server of the challenge. Models for this setting are trained with the union of training and validation set.

In experiments, we compare the performance of very deep two stream CNN [16] using several network architectures, including BN-Inception [4], Inception V3 [10], and

ResNet [3]. The performance of different network structures for spatial and temporal stream are summarized in Table 1. To analyze the effect of different training strategies, we compare the performance of appearance modeling CNNs with these strategies. The results are presented in Table 2. The contributions of appearance and motion CNNs are also summarized in Table 3. Then we report the performance of the two components in the acoustic analysis systems in Table 4.

Finally, we evaluate the fusion of visual analysis system and audio analysis system on both the validation and testing set. The results are illustrated in Table 5. The best mAP achieved by the final ensemble is 93.2%. We also took one chance on the testing server to evaluate a combination of one appearance CNN and one motion CNN. Its results are presented as "Visual CNN (Single)" in Table 5. It is exciting to see using this "single model" setting we can still achieve a reasonable mAP of 91.2%, which may better fit for industrial applications.

4. Conclusions

This paper has proposed an action recognition method for classifying temporally untrimmed videos. It is based on the idea of combining visual analysis and acoustic analysis. The results show that by carefully designing the visual and acoustic analysis systems and combining them, we can achieve exciting results in video classification tasks and boost the performance of state-of-the-art methods. Another fact to be noticed is that this high accuracy is achieved by evaluating only 1 frame per second, equivalent to only seeing around 4% of all frames of input videos. We believe this property is also very important for practically applying the system in industrial scenarios.

5. Acknowledgment

References

[1] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1, 3

[2] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015. 2

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 3

[4] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 3

[5] D. OShaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008. 2

[6] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109 – 125, 2016. 2

[7] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013. 2

[8] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 1

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2, 3

[11] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv preprint arXiv:1604.07160*, 2016. 2

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1

[13] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013. 1, 2

[14] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015. 1, 2

[15] L. Wang, Y. Qiao, and X. Tang. MoFAP: A multi-level representation for action recognition. *IJCV*, pages 1–18, 2015. 1

[16] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. 1, 2, 3

[17] Z. Wu, Y. Jiang, X. Wang, H. Ye, X. Xue, and J. Wang. Fusing multi-stream deep networks for video classification. *CoRR*, abs/1509.06086, 2015. 2

[18] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *CVPR*, pages 1600–1609, 2015. 2

[19] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *CVPR*, pages 2718–2726. 2

[20] Z. Zhu, J. H. Engel, and A. Hannun. Learning multi-scale features directly from waveforms. *arXiv preprint arXiv:1603.09509v2*, 2016. 3